Comparing the Medians of a Random Interval Defined by Means of Two Different L^1 Metrics

Beatriz Sinova[1] and Stefan Van Aelst[2]

Abstract The standard central tendency measure for interval-valued data is the Aumann-type expected value, but as in real settings it is not always convenient because of the big influence that small changes in the data as well as the existence of great magnitude data have on its estimate. The aim of this paper is to explore other summary measures with a more robust behavior. The real-valued case has served as inspiration to define the median of a random interval. The definition of the median as a 'middle position' value is not possible here because of the lack of a universally accepted total order in the space of interval data, so the median is defined as the element which minimizes the mean distance, in terms of an L^1 metric (extension of the Euclidean distance in $\mathbb{R}$), to the values the random interval can take. The two metrics that we consider are the generalized Hausdorff metric (like the well-known Hausdorff metric, but including a positive parameter which determines the relative importance given to the difference in imprecision with respect to the difference in location) and the 1-norm metric introduced by Vitale. The aim of this paper is to compare these two approaches for the median of a random interval, both theoretically based on concepts commonly used in robustness and empirically by simulation.

1 Introduction and Motivation

Statistical data obtained from random experiments can be of a very different nature. Interval data frequently appear when intrinsically imprecise measurements (like fluctuations, ranges, censoring times, etc.) or values associated with some imprecise knowledge on numerical values (when dealing

[1] Universidad de Oviedo, Calvo Sotelo s/n, 33007 Oviedo, Spain, sinovabeatriz@uniovi.es
[2] Universiteit Gent, Krijgslaan 281, S9, B-9000 Gent, Belgium, Stefan.VanAelst@UGent.be

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 75–86.
springerlink.com

with grouped data for instance) are involved. Many examples can be found in real life, such as the intervals describing the age range covered by each class when individuals in surveys are split into age groups, the fluctuation of quotations on the stock exchange or the temperature range for the daily forecast in a certain location. Similarly, many interval data sets are obtained in research studies in different fields such as Medicine, Engineering, Empirical and Social Sciences in which the information about the range of values the variable takes along a period is even more relevant than the detailed records.

Random intervals are interval-valued random elements, that is, they formalize mathematically the random mechanism of producing interval data associated with a random experiment. To analyze this type of data, some central tendency measures based on the interval arithmetic (globally considering intervals as elements and not as sets of elements) have been proposed. The most often used measure is the Aumann-type expected value. It inherits very good probabilistic and statistical properties from the mean of a real-valued random variable, but that is also the reason why it can be highly influenced by the existence of great magnitude data or data changes.

In real settings, the solution is to consider more robust central tendency measures, like the median. Inspired by this, we define the median of a random interval. Taking into account that there is no universally accepted total order criterion in the space of non-empty compact intervals (so the median cannot be defined as a 'middle' position value), an L^1 metric, generalization of the Euclidean metric in $\mathbb{R}$, is required to define the median as the element of the space minimizing the mean distance to all the values the random interval can take. The first choice for the L^1 metric was the generalized Hausdorff metric (see Sinova *et al* [5]): a new distance based on the well-known Hausdorff metric expressed in terms of the mid/spr characterization of intervals (that is, their mid-point and their spread or radius). However, there are obstacles to generalize the median defined by means of the generalized Hausdorff metric to random fuzzy numbers due to the fact that, although the generalized mid and spread (see Trutschnig *et al.* [7]) characterize a fuzzy number, the sufficient and necessary conditions a function must fulfill to be a generalized mid or spread are not known yet and it is not possible to guarantee that the median defined in that way is indeed a fuzzy number. These difficulties prompted the use of another distance (suitable for the definition of the median of a random fuzzy number as shown in Sinova *et al.* [6]), based on the 1-norm, as introduced by Vitale [8], and which considers the characterization of an interval in terms of infima and suprema. Of course, a second definition of median of random intervals is obtained as a particular case of the median for random fuzzy numbers. The definition of both medians and their immediate properties are studied in Section 3, after recalling in Section 2 the notation and basic operations and concepts in the space of interval data. In Section 4, the two proposed definitions of median of a random interval are compared by means of the finite sample breakdown point and some simulation studies. Finally, Section 5 presents some conclusions and open problems.

2 The Space of Intervals $\mathcal{K}_c(\mathbb{R})$: Preliminaries

First of all, some notation is established, starting with $\mathcal{K}_c(\mathbb{R})$, the class of nonempty compact intervals. Each one of the intervals $K \in \mathcal{K}_c(\mathbb{R})$ can be characterized in terms of its infimum and supremum, $K = [\inf K, \sup K]$ or in terms of its mid-point and spread or radius, $K = [\operatorname{mid} K - \operatorname{spr} K, \operatorname{mid} K + \operatorname{spr} K]$, where

$$\operatorname{mid} K = \frac{\inf K + \sup K}{2}, \quad \operatorname{spr} K = \frac{\sup K - \inf K}{2}.$$

To analyze this kind of data, the two most relevant operations from a statistical point of view are the addition and the product by a scalar. In this paper, we use the usual interval arithmetic (the particular case of set arithmetic). That is:

- The *sum* of two nonempty compact intervals, $K, K' \in \mathcal{K}_c(\mathbb{R})$, is defined as the Minkowski sum of K and K', i.e., as the interval

$$K + K' = [\inf K + \inf K', \sup K + \sup K'] =$$

$$[(\operatorname{mid} K + \operatorname{mid} K') - (\operatorname{spr} K + \operatorname{spr} K'), (\operatorname{mid} K + \operatorname{mid} K') + (\operatorname{spr} K + \operatorname{spr} K')].$$

- The *product of an interval* $K \in \mathcal{K}_c(\mathbb{R})$ *by a scalar* $\gamma \in \mathbb{R}$ is defined as the element of $\mathcal{K}_c(\mathbb{R})$ such that

$$\gamma \cdot K = \begin{cases} [\gamma \cdot \inf K, \gamma \cdot \sup K] \text{ if } \gamma \geq 0 \\ [\gamma \cdot \sup K, \gamma \cdot \inf K] \text{ otherwise} \end{cases}$$

$$= [\gamma \cdot \operatorname{mid} K - |\gamma| \cdot \operatorname{spr} K, \gamma \cdot \operatorname{mid} K + |\gamma| \cdot \operatorname{spr} K].$$

A very important remark is that with these two operations the space is not linear, but only semilinear (with a conical structure) because of the lack of an opposite element for the Minkowski addition. Therefore, there is no generally applicable definition for the difference of intervals that preserves the connection with the sum in the real case. Hence, distances play a crucial role in statistical developments. Although L^2 metrics are very convenient in many statistical developments like least squares approaches, an L^1 distance is now needed in order to define the median. In this paper, the two following L^1 metrics will be used:

- The *generalized Hausdorff metric* (Sinova *et al.* [5]), which is partially inspired by the Hausdorff metric for intervals and the L^2 metrics in Trutschnig *et al.* [7]. It includes a positive parameter to weight the relative importance of the distance between the spreads relative to the distance between the mid-points (allocating the same weight to the deviation in location as to the deviation in imprecision is often viewed as a concern in the Hausdorff metric). Given two intervals $K, K' \in \mathcal{K}_c(\mathbb{R})$ and any

$\theta \in (0, \infty)$, the generalized Hausdorff metric between them is defined as:

$$d_{H,\theta}(K, K') = |\text{mid}\, K - \text{mid}\, K'| + \theta \cdot |\text{spr}\, K - \text{spr}\, K'|.$$

- The *1-norm metric*, introduced by Vitale [8]. Given any two intervals $K, K' \in \mathcal{K}_c(\mathbb{R})$, the 1-norm distance between them is:

$$\rho_1(K, K') = \frac{1}{2}|\inf K - \inf K'| + \frac{1}{2}|\sup K - \sup K'|.$$

  As mentioned before, this corresponds to the particular case (for intervals) of the metric used to define the median of random fuzzy numbers (Sinova *et al.* [5]).

A *random interval* is usually defined (following the random set-based approach to introduce this notion) as a Borel measurable mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$, starting from a probability space $(\Omega, \mathcal{A}, P)$, with respect to $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the Hausdorff metric. The generalized Hausdorff metric and the 1-norm metric are topologically equivalent to each other and to the Hausdorff metric. Therefore, the definition of random interval can be rewritten in terms of either of these two metrics instead of the Hausdorff metric. This Borel measurability guarantees that concepts like the *distribution induced by a random interval* or the *stochastic independence of random intervals*, crucial for inferential developments, are well-defined by trivial induction. A random interval can also be defined in terms of real-valued random variables: X is a random interval if, and only if, both functions $\inf X : \Omega \to \mathcal{K}_c(\mathbb{R})$ and $\sup X : \Omega \to \mathcal{K}_c(\mathbb{R})$ (or equivalently, $\text{mid}\, X : \Omega \to \mathcal{K}_c(\mathbb{R})$ and $\text{spr}\, X : \Omega \to [0, \infty)$) are real-valued random variables.

The *Aumann expectation* is the standard central tendency measure for random intervals. This mean value is indeed the Fréchet expectation with respect to the d_θ metric, which corresponds to the Bertoluzza *et al.* [1] distance (see Gil *et al.* [3]) for the particular case of interval-valued data, and is defined as:

$$d_\theta(K, K') = \sqrt{(\text{mid}\, K - \text{mid}\, K')^2 + \theta \cdot (\text{spr}\, K - \text{spr}\, K')^2},$$

where $K, K' \in \mathcal{K}_c(\mathbb{R})$ and $\theta \in (0, \infty)$. This means that the Aumann expectation is the unique interval which minimizes, over $K \in \mathcal{K}_c(\mathbb{R})$, the expected squared distance $E[(d_\theta(X, K))^2]$. Furthermore, it can be expressed explicitly as the interval whose mid-point equals the expected value of $\text{mid}\, X$ and whose spread equals the expected value of $\text{spr}\, X$. The Aumann expectation inherits many very good probabilistic and statistical properties from the expectation of a real-valued random variable, like the linearity and invariance under linear transformations, and it also fulfills the Strong Law of Large Numbers for almost all the metrics we can consider. However, its high sensitivity to

data changes or extreme data makes this value not always convenient when summarizing the information given by interval-valued data sets.

3 The Median of a Random Interval Defined Through an L^1 Metric

The Aumann expectation of a random interval is not robust enough which is the motivation for extending the concept of median. Nevertheless, the non-existence of a universally accepted total order in the space $\mathcal{K}_c(\mathbb{R})$ does not allow us to define it as a 'middle position' value. In real settings another approach is to define the median as the value with the smallest mean Euclidean distance to the values of the real-valued random variable. Then, an L^1 metric between intervals which extends the Euclidean distance is required in order to define the median as the interval with the smallest mean distance to the values of the random interval. The two L^1 metrics between intervals introduced before satisfy this condition, so the definition of the median through both distances is now formalized.

Definition 1. The $d_{H,\theta}$-median (or medians) of a random interval $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is (are) defined as the interval(s) $\mathrm{Me}[X] \in \mathcal{K}_c(\mathbb{R})$ such that:

$$E(d_{H,\theta}(X, \mathrm{Me}[X])) = \min_{K \in \mathcal{K}_c(\mathbb{R})} E(d_{H,\theta}(X, K)), \tag{1}$$

if these expected values exist.

A very practical result that guarantees the existence of the median and allows to compute it is the following. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated random interval X, the minimization problem (1) has at least one solution, given by any nonempty compact interval such that:

$$\mathrm{mid}\ \mathrm{Me}[X] = \mathrm{Me}(\mathrm{mid}\, X), \quad \mathrm{spr}\ \mathrm{Me}[X] = \mathrm{Me}(\mathrm{spr}\, X).$$

It can immediately be noticed that the $d_{H,\theta}$-median is not unique if either $\mathrm{Me}(\mathrm{mid}\, X)$ or $\mathrm{Me}(\mathrm{spr}\, X)$ (which are medians of real-valued random variables) are not unique. It should be pointed out that the chosen solution does not depend on the value chosen for theta, although the mean error does.

Analogously, the median can be defined by means of the 1-norm metric:

Definition 2. The ρ_1-median (or medians) of a random interval $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is (are) defined as the interval(s) $\mathrm{Med}[X] \in \mathcal{K}_c(\mathbb{R})$ such that:

$$E(\rho_1(X, \mathrm{Med}[X])) = \min_{K \in \mathcal{K}_c(\mathbb{R})} E(\rho_1(X, K)), \tag{2}$$

if these expected values exist.

In this situation, the practical choice (one of the solutions of minimization problem (2)) is the interval $\mathrm{Med}[X] \in \mathcal{K}_c(\mathbb{R})$ which satisfies:

$$\inf \mathrm{Med}[X] = \mathrm{Me}(\inf X), \quad \sup \mathrm{Med}[X] = \mathrm{Me}(\sup X).$$

If any of these two medians of real-valued random variables are not unique, the usual criterion of choosing the mid-point of the interval of possible medians is used to guarantee that $\mathrm{Med}(X)$ is nonempty.

Both medians preserve most of the elementary operational properties of the median in real settings. Namely,

Proposition 1. *Suppose that X is a random interval associated with a probability space. Then,*

- *if the distribution of X is degenerate at an interval value $K \in \mathcal{K}_c(\mathbb{R})$,*

$$\mathrm{Me}[X] = K,$$
$$\mathrm{Med}[X] = K.$$

- *for any $K \in \mathcal{K}_c(\mathbb{R})$ and $\gamma \in \mathbb{R}$,*

$$\mathrm{Me}[\gamma \cdot X + K] = \gamma \cdot \mathrm{Me}[X] + K,$$
$$\mathrm{Med}[\gamma \cdot X + K] = \gamma \cdot \mathrm{Med}[X] + K.$$

One remark about a distinctive feature in contrast to the real-valued case is that neither the $d_{H,\theta}$-median nor the ρ_1-median of a random interval is necessarily a value taken by the random interval as can be noticed from the following example: let X be a random interval taking the values $[0, 4]$, $[1, 3]$ and $[2, 5]$ with probability $\frac{1}{3}$. In this situation, the $d_{H,\theta}$-median is the interval $\mathrm{Me}[X] = [\mathrm{Me}(\mathrm{mid}\, X) - \mathrm{Me}(\mathrm{spr}\, X), \mathrm{Me}(\mathrm{mid}\, X) + \mathrm{Me}(\mathrm{spr}\, X)] = [2 - \frac{3}{2}, 2 + \frac{3}{2}] = [\frac{1}{2}, \frac{7}{2}]$ and the ρ_1-median is $\mathrm{Med}[X] = [\mathrm{Me}(\inf X), \mathrm{Me}(\sup X)] = [1, 4]$, neither of them being values the random interval takes.

As mentioned before, there is no universally accepted total order in the space $\mathcal{K}_c(\mathbb{R})$, so it is not possible to define the median as a 'middle position' value. However, both medians are a *measure of 'middle position'* with a certain partial ordering, when applicable. For the $d_{H,\theta}$-median, it can be proven that it is coherent with the Ishibuchi and Tanaka [4] partial ordering:

$$K \leq_{CW} K' \text{ if, and only if, } \mathrm{mid}\, K \leq \mathrm{mid}\, K' \text{ and } \mathrm{spr}\, K \geq \mathrm{spr}\, K'.$$

Hence, K' is considered to be *CW-larger* than K if, and only if, its location is greater and its imprecision is lower than for K:

Proposition 2. *For any sample of individuals $(\omega_1, \ldots, \omega_n)$ such that*

$$X(\omega_1) \leq_{CW} \ldots \leq_{CW} X(\omega_n)$$

we have that

- *if n is an odd number, then* $\mathrm{Me}[X] = X(\omega_{(n+1)/2})$,

- *if n is an even number, then* $\mathrm{Me}[X]$ = *any interval value 'between'* $X(\omega_{n/2})$ *and* $X(\omega_{(n/2)+1})$*, the 'between' being intended in the* $\leq_{CW}$ *sense, that is,* $\mathrm{mid}\, Me[X]$ *can be any value in* $\left[\mathrm{mid}\, X(\omega_{n/2}), \mathrm{mid}\, X(\omega_{(n/2)+1})\right]$*, whereas* $\mathrm{spr}\, Me[X]$ *can be any value in* $\left[\mathrm{spr}\, X(\omega_{(n/2)+1}), \mathrm{spr}\, (\omega_{n/2})\right]$.

On the other hand, the ρ_1-median is coherent with the well-known product order for the inf/sup vector, which is the partial ordering given by:

$$K \precsim K' \text{ if, and only if, } \inf K \leq \inf K' \text{ and } \sup K \geq \sup K'$$

or, equivalently, for all $\lambda \in [0,1]$ we have that $K^{[\lambda]} \leq K'^{[\lambda]}$, where $K^{[\lambda]} = \lambda \sup K + (1-\lambda) \inf K$.

Proposition 3. *For any sample of individuals* $(\omega_1, \ldots, \omega_n)$ *such that*

$$X(\omega_1) \precsim \ldots \precsim X(\omega_n)$$

we have that

- *if n is an odd number, then* $\mathrm{Med}[X] = X(\omega_{(n+1)/2})$,
- *if n is an even number, then* $\mathrm{Med}[X] = \dfrac{X(\omega_{n/2}) + X(\omega_{(n/2)+1})}{2}$.

Finally, the strong consistency of both the sample $d_{H,\theta}$-median and the sample ρ_1-median as estimators of the corresponding population quantities can be proven under very mild conditions as shown in the following results.

Proposition 4. *Suppose that X is a random interval associated with a probability space* $(\Omega, \mathcal{A}, P)$ *and* $\mathrm{Me}[X]$ *is unique. If* $\widehat{\mathrm{Me}[X]}_n$ *denotes the sample median associated with a simple random sample* $(X_1, \ldots, X_n)$ *from X, then*

$$\lim_{n\to\infty} d_{H,\theta}\left(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[X]\right) = 0 \quad a.s.[P].$$

Proposition 5. *Suppose that X is a random interval associated with a probability space* $(\Omega, \mathcal{A}, P)$ *and* $\mathrm{Med}[X]$ *is unique without applying any convention. If* $\widehat{\mathrm{Med}[X]}_n$ *denotes the sample median associated with a simple random sample* $(X_1, \ldots, X_n)$ *from X, then*

$$\lim_{n\to\infty} \rho_1\left(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[X]\right) = 0 \quad a.s.[P].$$

4 The Comparison between the $d_{H,\theta}$-median and the ρ_1-median of a Random Interval

The first result compares the $d_{H,\theta}$-median and the ρ_1-median by means of the computation of the finite sample breakdown point. Recall that the finite sample breakdown point is a measure of the robustness, since it gives the minimum proportion of sample data which should be arbitrarily increased or

decreased to make the estimate arbitrarily large or small. Following Donoho and Huber [2], the *finite sample breakdown point* (fsbp) of the sample $d_{H,\theta}$-median in a sample of size n from a random interval X is given by:

$$\text{fsbp}(\widehat{\text{Me}[X]}_n, x_n, d_{H,\theta})$$

$$= \frac{1}{n} \min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} d_{H,\theta}(\widehat{\text{Me}[X]}_n, \widehat{\text{Me}[Y_k]}_n) = \infty \right\},$$

where x_n denotes the considered sample of n data from the metric space $(\mathcal{K}_c(\mathbb{R}), d_{H,\theta})$ in which $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} d_{H,\theta}(K, K') = \infty$ and $\text{Me}[Y_k]_n$ is the sample median of the sample $y_{n,k}$ obtained from the original sample x_n by perturbing at most k observations.

Analogously, the finite sample breakdown point of the sample ρ_1-median in a sample of size n from a random interval X is, with the same notation:

$$\text{fsbp}(\widehat{\text{Med}[X]}_n, x_n, \rho_1)$$

$$= \frac{1}{n} \min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} \rho_1(\widehat{\text{Med}[X]}_n, \widehat{\text{Med}[Y_k]}_n) = \infty \right\},$$

Then, it can be proven that

Proposition 6. *The finite sample breakdown point of both the sample $d_{H,\theta}$-median and the ρ_1-median from a random interval X, equal*

$$\text{fsbp}(\widehat{\text{Me}[X]}_n, x_n, d_{H,\theta}) = \text{fsbp}(\widehat{\text{Med}[X]}_n, x_n, \rho_1) = \frac{1}{n} \cdot \lfloor \frac{n+1}{2} \rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Proof. First note that the conditions $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} d_{H,\theta}(K, K') = \infty$ and $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} \rho_1(K, K') = \infty$ are fulfilled in the corresponding metric spaces because $d_{H,\theta}(\mathbf{1}_{[n-1,n+1]}, \mathbf{1}_{[-n-1,-n+1]}) = \rho_1(\mathbf{1}_{[n-1,n+1]}, \mathbf{1}_{[-n-1,-n+1]}) = 2n$. Since the fsbp for the sample median of a real-valued random variable equals $\lfloor \frac{n+1}{2} \rfloor$, we immediately have that:

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\text{mid}\,[X]}_n) - \text{Me}(\widehat{\text{mid}\,[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\text{spr}\,[X]}_n) - \text{Me}(\widehat{\text{spr}\,[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\inf[X]}_n) - \text{Me}(\widehat{\inf[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\sup[X]}_n) - \text{Me}(\widehat{\sup[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

Therefore,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} d_{H,\theta}(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n) \geq$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{mid}\,(\widehat{\mathrm{Me}[X]}_n) - \mathrm{mid}\,(\mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n)| =$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{Me}(\widehat{\mathrm{mid}[\mathrm{X}]}_\mathrm{n}) - \mathrm{Me}(\widehat{\mathrm{mid}[\mathrm{Y}_{\lfloor\frac{n+1}{2}\rfloor}]}_\mathrm{n})| = \infty$$

and

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} \rho_1(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n) \geq$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} \frac{1}{2}|\inf(\widehat{\mathrm{Med}[X]}_n) - \inf(\mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n)|$$

$$= \frac{1}{2} \sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor}]}_n)| = \infty$$

On the other hand,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\mathrm{mid}\,[X]}_n) - \mathrm{Me}(\mathrm{mid}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_1 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\mathrm{spr}\,[X]}_n) - \mathrm{Me}(\mathrm{spr}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_2 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_3 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\sup[X]}_n) - \mathrm{Me}(\widehat{\sup[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_4 < \infty$$

Consequently,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} d_{H,\theta}(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor-1}}]_n)$$

$$= \sup_{Y_{\lfloor\frac{n+1}{2}\rfloor-1}} \Big[|\mathrm{Me}(\widehat{\mathrm{mid}[X]}_n) - \mathrm{Me}(\mathrm{mid}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|$$

$$+ \theta \cdot |\mathrm{Me}(\widehat{\mathrm{spr}[X]}_n) - \mathrm{Me}(\mathrm{spr}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|\Big] \leq M_1 + \theta \cdot M_2 < \infty$$

and

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \rho_1(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)$$

$$= \sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \left[\frac{1}{2} \cdot |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|\right.$$

$$+\frac{1}{2}\cdot\left|\mathrm{Me}(\widehat{\sup[X]}_n)-\mathrm{Me}(\widehat{\sup[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)\right|\Big]\leq\frac{M_3+M_4}{2}<\infty$$

□

Furthermore, the fsbp of both medians can also be compared with the Aumann expectation:

Theorem 1. *The finite sample breakdown point of the sample Aumann expectation from a random interval* X, $\mathrm{fsbp}(\overline{X_n})$, *is lower than the ones for the sample* $d_{H,\theta}$*-median and the sample* ρ_1*-median for samples of size* $n>2$.

Proof. Following the same reasoning used in the previous proposition, it can be proven that

$$\mathrm{fsbp}(\overline{X_n},x_n,d_{H,\theta})=\mathrm{fsbp}(\overline{X_n},x_n,\rho_1)=\frac{1}{n},$$

so, consequently,

$$\mathrm{fsbp}(\widehat{\mathrm{Me}[X]}_n,x_n,d_{H,\theta})\geq\frac{n/2}{n}=\frac{1}{2}>\frac{1}{n}=\mathrm{fsbp}(\overline{X_n},x_n,d_{H,\theta})$$

$$\mathrm{fsbp}(\widehat{\mathrm{Med}[X]}_n,x_n,\rho_1)\geq\frac{n/2}{n}=\frac{1}{2}>\frac{1}{n}=\mathrm{fsbp}(\overline{X_n},x_n,\rho_1)$$

□

In order to corroborate these results, some empirical studies have been developed. A sample of $n=10000$ interval-valued data has been randomly generated from a random interval characterized by the distribution of two real-valued random variables, $\mathrm{mid}\,X$ and $\mathrm{spr}\,X$. Two cases have been considered: one in which the two random variables are independent (Case 1) and another one in which they are dependent (Case 2). In both situations, the sample has been split into two subsamples, one of size $n\cdot c_p$ associated with a contaminated distribution (hence c_p represents the proportion of contamination) and the other one, of size $n\cdot(1-c_p)$, without any perturbation. A second parameter, C_D, has also been included to measure the relative distance between the distribution of the contaminated and non contaminated subsamples. In detail, for different values of c_p and C_D the data for Case 1 are generated according to

- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,1)$ and $\mathrm{spr}\,X\rightsquigarrow\chi^2_1$ for the non contaminated subsample,
- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,3)+C_D$ and $\mathrm{spr}\,X\rightsquigarrow\chi^2_4+C_D$ for the contaminated subsample,

while for Case 2 we use

- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,1)$ and $\mathrm{spr}\,X\rightsquigarrow\left(\frac{1}{(\mathrm{mid}\,X)^2+1}\right)^2+.1\cdot\chi^2_1$ for the non contaminated subsample,

- $\operatorname{mid} X \rightsquigarrow \mathcal{N}(0,3) + C_D$ and $\operatorname{spr} X \rightsquigarrow \left(\frac{1}{(\operatorname{mid} X)^2+1}\right)^2 + .1 \cdot \chi_1^2 + C_D$ for the contaminated subsample.

Both the population $d_{H,\theta}$-median and the population ρ_1-median are approximated by the Monte Carlo approach from this sample and the expected distance between the non contaminated distribution, X_{nc}, and the approximated medians, considering the $d_{H,\theta}$ and the ρ_1 distances, were computed.

c_p	c_D	Ratio_ρ	$\text{Ratio}_{\theta=1/3}$	$\text{Ratio}_{\theta=\sqrt{1/3}}$	$\text{Ratio}_{\theta=1}$	Ratio_ρ	$\text{Ratio}_{\theta=1/3}$	$\text{Ratio}_{\theta=\sqrt{1/3}}$	$\text{Ratio}_{\theta=1}$
.0	0	1.019406	1.010211	1.014805	1.020016	1.090363	1.071163	1.113693	1.173596
.0	1	1.019391	1.010212	1.014806	1.020017	1.090412	1.071170	1.113704	1.173612
.0	5	1.019393	1.010221	1.014805	1.020014	1.090448	1.071139	1.113654	1.173533
.0	10	1.019410	1.010209	1.014802	1.020012	1.090442	1.071171	1.113705	1.173613
.1	0	1.016934	1.008394	1.012000	1.015977	1.081155	1.066063	1.106141	1.163368
.1	1	1.017550	1.008663	1.012288	1.016226	1.072844	1.053439	1.085163	1.129607
.1	5	1.015010	1.007975	1.011077	1.014365	1.065343	1.046932	1.071485	1.102874
.1	10	1.011805	1.006462	1.008901	1.011478	1.046427	1.036585	1.054048	1.075179
.2	0	1.014011	1.006560	1.009286	1.012245	1.073723	1.061916	1.099925	1.154805
.2	1	1.014893	1.006741	1.009424	1.012272	1.056341	1.037449	1.059556	1.090469
.2	5	1.012616	1.006605	1.008862	1.011194	1.047532	1.028887	1.041835	1.057560
.2	10	1.009017	1.004951	1.006547	1.008209	1.029309	1.020252	1.028132	1.037146
.4	0	1.008012	1.003628	1.005115	1.006738	1.062304	1.055413	1.090023	1.140840
.4	1	1.006726	1.003075	1.004202	1.005429	1.024528	1.014247	1.022384	1.033863
.4	5	1.009291	1.007752	1.008795	1.009980	1.022742	1.014213	1.018332	1.022988
.4	10	1.006831	1.007307	1.008008	1.008899	1.012734	1.009485	1.011648	1.013964
.4	100	1.000904	1.000999	1.001095	1.001233	1.001385	1.001161	1.001371	1.001585

Table 1. Ratios of the mean distances of the mixed (partially contaminated and non-contaminated) sample $d_{H,\theta}$ and ρ_1-medians to the non-contaminated distribution of a random interval in Case 1 (left columns) and Case 2 (right columns)

In Table 1, the ratios $\text{Ratio}_\rho = E(\rho_1(X_{nc}, \text{Me}[X]))/E(\rho_1(X_{nc}, \text{Med}[X]))$ and $\text{Ratio}_\theta = E(d_{H,\theta}(X_{nc}, \text{Med}[X]))/E(d_{H,\theta}(X_{nc}, \text{Me}[X]))$ are shown. They show us how the mean distance increases (w.r.t. each metric) when the chosen median is not the one defined by means of the corresponding metric.

As Table 1 shows, the bigger the error proportion, the smaller the ratios. It can be also noticed that the smaller the θ, the smaller the corresponding ratio. As all the ratios are very close to 1, it can be concluded that both the $d_{H,\theta}$-median (with different choices for θ) and ρ_1-median have a quite similar behavior since there are no big differences between choosing one of the two measures in order to summarize the information given by the sample (independently from the distance used).

5 Concluding Remarks

In this study, two different definitions for the median of a random interval have been compared. Both definitions preserve important properties of the median in real settings and are coherent with the interpretation of the median as a 'middle position' value for certain partial orderings between intervals. By calculating the finite sample breakdown point and some simulation studies, the robustness of the two medians has been shown to be similar.

Future directions to be considered could be the extension of this comparison to the fuzzy-valued case and the definition of other central tendency measures. For instance, trimmed means or medians defined through depth functions could be adapted to this situation and compared with the current results.

Acknowledgements The research by Beatriz Sinova was partially supported by / benefited from the Spanish Ministry of Science and Innovation Grant MTM2009-09440-C02-01 and the COST Action IC0702. She has been also granted with the Ayuda del Programa de FPU AP2009-1197 from the Spanish Ministry of Education, an Ayuda de Investigación 2011 from the Fundación Banco Herrero and three Short Term Scientific Missions associated with the COST Action IC0702. The research by Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen). Their financial support is gratefully acknowledged.

References

1. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware & Soft Comput.* 2:71–84
2. Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum K, Hodges Jr. JL (eds.) *A Festschrift for Erich L. Lehmann.* Wadsworth, Belmont
3. Gil MA, Lubiano MA, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval data. *Metrika* 56:97–111
4. Ishibuchi H, Tanaka H (1990) Multiobjective programming in optimization of the interval objective function. *Europ. J. Oper. Res.* 48:219–225
5. Sinova B, Casals MR, Colubi A, Gil MA (2010) The median of a random interval. In: Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O (eds.) *Combining Soft Computing and Statistical Methods in Data Analysis.* Springer, Heidelberg
6. Sinova B, Gil MA, Colubi A, Van Aelst S (2011) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.* in press (doi:10.1016/j.fss.2011.11.004)
7. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inf. Sci.* 179:3964–3972
8. Vitale RA (1985) L_p metrics for compact, convex sets. *J. Approx. Theory* 45:280–287